Advancing Audio-Based Text Generation with Imbalanced Preference Optimization

Zhenghao Zhou[1], **Yongjie Liu**[1], **Chen Cao**[2*]

[1] National Supercomputing Center in Wuxi, China
[2] University of Sheffield, United Kingdom
zhenghaozhou183@gmail.com, ccao5@Sheffield.ac.uk

Abstract

Human feedback in generative systems is a highly active frontier of research that aims to improve the quality of generated content and align it with subjective preferences. Existing efforts predominantly focus on text-only large language models (LLMs) or text-based image generation, while cross-modal generation between audio and text remains largely unexplored. Moreover, there is currently no open-source preference dataset to support the deployment of alignment algorithms in this domain. In this work, we take audio speech translation (AST) and audio captioning (AAC) tasks as examples to explore how to enhance the performance of mainstream audio-based text generation models with limited human feedback. Specifically, we propose an novel framework named imbalanced preference optimization (IPO) that includes a model adversarial sampling concept–human annotators act as referees to determine model outcomes, using these results as pseudo-labels for the corresponding beam search hypotheses. Given these imbalanced win-loss results, IPO effectively enable the two models to update interactively to win the next round of adversarial sampling. We conduct both subjective and objective evaluations to demonstrate the alignment benefits of IPO and its enhancement on model perception and generation capacities. On both AAC and AST, a few hundreds of annotations significantly enhance the weak model, and the strong model can also be encouraged to achieve new state-of-the-art results in terms of objective metrics. Additionally, we show the extensibility of IPO by applying it to the reverse task of text-to-speech generation, improving system robustness on unseen reference speaker.

1 Introduction

In the wave of AI-Generated Content (AIGC), learning from human feedback plays a crucial role that aims to align the generated content of AI systems with human preference (MacGlashan et al. 2017; Stiennon et al. 2020; Dubois et al. 2024). For example, large language models (LLMs) undergo calibration through Reinforcement Learning from Human Feedback (RLHF) to become helpful and powerful assistant systems like ChatGPT (Bai et al. 2022; Achiam et al. 2023). Additionally, recent efforts have demonstrated the efficacy of RLHF in high-dimensional data generation tasks, e.g., music generation (Cideron et al. 2024) and text-to-images synthesis (Wu et al. 2023; Liang et al. 2024).

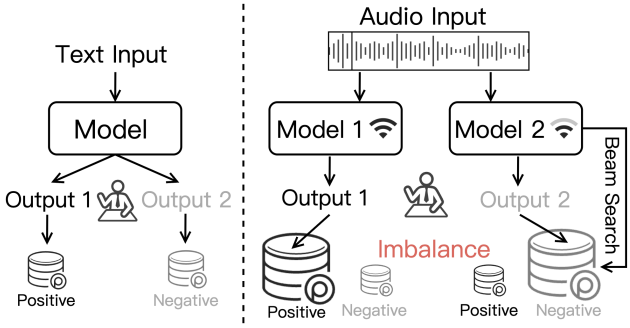

Figure 1: *Left:* text-based preference annotation, where language model generates desired/undesired samples annotated by human. *Right:* audio-text adversarial annotations. Due to lack of diversity, annotator serves as referee to determine the win-loss results of two models based on the same audio input. Additionally, Beam search algorithm is utilized for extending data with pseudo-preference.

In addition to preference alignment, RLHF-based methods have also been reported in existing works to optimize model performance evaluated by objective evaluation metric (Liu et al. 2020; Steenhoek et al. 2023), particularly in autoregressive sequence prediction (Chen et al. 2024a). A underlying reason is that the sampling process in RLHF can effectively mitigate the exposure bias problem caused by the mismatch between teacher-forcing training and autoregressive inference (Rennie et al. 2017). Meanwhile, these self-generated samples, after annotated by humans, can provide high-level supervised information to the neural model through gradient descent. This is certainly beneficial for tasks where the ground truth is not unique, such as machine translation (Xu et al. 2024; He et al. 2024) and image captioning (Xu et al. 2023; Verma et al. 2023).

Unlike aforementioned text-only and vision-based generation tasks, which have widely optimized by RLHF, we find that there is a considerable research gap in audio-based generation tasks when predicting the natural language consuming a segment of audio signal or human speech. Specifically, in relative speech translation and audio captioning tasks, few prior work incorporates subjective feedback into the learn-

*Corresponding author

ing process of neural networks (Chu et al. 2024). Motivated by this, we focus our research on addressing this gap and are the first to investigate the following questions: Can RLHF align audio-text cross-modal generation systems with human preferences, or can it improve model performance by objective metrics?

In this context, the primary challenge is the lack of labeled preference data. Compared to image or text data annotation, annotating audio is more labor-intensive (Gibbons et al. 2023): participants can straightforward perceive images or text, but for audio, they need to listen to the entire playback, even listening several times to prevent forgetting. Consequently, the limited accessible data imposes higher demands on data efficiency for audio-text alignment. Furthermore, the audio modality exhibits complex and variable characteristics across domains, necessitating a model that can not only generate natural language but also possess robust audio pattern recognition capabilities. For example, if the speech translation model misrecognizes homophones of input, it can explicitly affect the quality of predicted text.

In this work, we propose a IPO, an preference optimization approach tailored to audio-text modality, achieving cross-modal alignment with limited subjective human feedback. In general, IPO is novel for two aspects: (i) To maximize the utility of human-annotated data, we introduce a model adversarial sampling: given an audio input, the annotators sever as a referee to determine the superiority of the outputs generated by two models. This judgment of superiority not only applies to the given input-output pair, but also extends to other hypotheses sampled by these two models through beam search decoding. Notably, we conduct empirical validation to confirm the rationality of this operation, which expand the self-generated negative samples to n times with a beam size of n. (ii) Considering that the two models may differ in strength, the subsequent optimization method is designed to effectively handle the data imbalance between positive and negative samples. Specifically, IPO does not depend on pairwise preference data, such as that used in DPO (Rafailov et al. 2024). Instead, it simply requires results of which model wins or loses in the adversarial sampling for the given data, then two models are interactively optimized to win the next round of confrontation.

To demonstrate the efficacy of IPO, we conduct intensive experiments based on prevalent audio-language foundation models with limited human annotations. Both objective and subjective evaluation reveal that IPO not only aligns with human feedback but also effectively enhances these model performance in terms of various objective metrics. (i) In AST task, only 100 data annotations can significantly enhance the weak model (Whisper-large-v2 (Radford et al. 2023)), more importantly, the performance of a stronger model (Seamless-v2 (Barrault et al. 2023)) can also be improved to achieve a new state-of-the-art according to the BLEU score. (ii) In AAC task, similar performance gain is observed. Moreover, we found that IPO can enhance the model's audio perception capabilities–some audio events that were previously unrecognized are now detected and included in the output descriptions. (iii) We also demonstrate the scalability of IPO by extending IPO to speech synthesis task to enhance the zero-shot capacity of mainstream text-to-speech (TTS) systems.

In general, our contributions are summarized as follows:

- We direct our focus on the alignment issue in audio-based text generation. In this context, human feedback is efficiently introduced into the learning loop of prevalent large audio models to further improve their performance.
- An RLHF-based iterative method, IPO, has been proposed to tackle the preference data scarcity. It introduces a model adversarial sampling and utilizes human annotators as referees, extending this win-loss result at a given data point to other self-generated responses through beam search. Subsequently, the IPO accordingly optimizes two models based on the unpaired and imbalanced preference data.
- Both subjective and objective evaluation demonstrate that IPO can effectively benefit to AST and AAC tasks, in terms of human preference and objective metric. Furthermore, we adapt IPO to the inverse text-based speech generation (i.e., TTS) task, improving its zero-shot robustness on cloning unseen speaker.

2 Related Work

2.1 Audio-Text Generation Models

With the advent of deep neural networks, the increasing availability of data has led to a surge in research utilizing both audio and text modalities for training neural networks (Rubenstein et al. 2023; Li, Tang, and Liu 2024), benefiting a multitude of cross-modal downstream tasks (Wang et al. 2024). In the realm of speech processing, extensively explored tasks include automatic speech recognition (Yang et al. 2024), automatic speech translation (Chen et al. 2024b), speech question-answering (Zhao et al. 2024), and spoken dialogue systems (Mousavi et al. 2024). From audio to speech, tasks such as audio captioning (Deshmukh et al. 2023) and audio question-answering (Chuang et al. 2019) have also seen significant advancements. More recently, all-in-one large models capable of handling both speech and audio have emerged, successfully executing text generation tasks based on varying instructions (Chu et al. 2023; Hu et al. 2024; Sun et al. 2024; Fathullah et al. 2024).

Moreover, the text-based audio generation task has also witnessed remarkable progress in recent years. Text-to-speech models now produce high-fidelity human-like speech and demonstrate the ability to generalize to unseen speakers in zero-shot scenarios (Kaur and Singh 2023; Wang et al. 2023; Peng et al. 2024). Additionally, text-to-audio (Huang et al. 2023) and text-to-music (Schneider et al. 2023; Kim, Jang, and Shin 2022) tasks have exhibited rich and high-quality synthesis outcomes.

2.2 Preference Optimization

The alignment of generated content with human preferences has been recognized as crucial since the early days of machine translation tasks. Following the recent successes of large language models, reinforcement learning from human

feedback (RLHF) has rapidly evolved to ensure that generated content is useful, non-toxic, and aligned with mainstream values.

Early RLHF approaches like PPO rely on an independent reward models trained by human-annotated data. However, the hacking arising from the reward model's inability to cover all scenarios. The methods directly optimizing with preference data becomes popular, such as DPO and its derivative works. Additionally, some methods have employed self-reward mechanisms, where the LLM itself is utilized to assign rewards, or LLM and the reward model update alternatively via a min-max game (Cheng et al. 2023). From a task perspective, RLHF has expanded beyond textual generation to include the generation of high-dimensional data such as images (Kirstain et al. 2023; Liang et al. 2024), videos (Zhou et al. 2024), audio (Majumder et al. 2024), music (Cideron et al. 2024), and speech (Chen et al. 2024a).

3 Methodology

Given a high-sample-rate audio input $x \in \mathbb{R}^T$, the audio-to-text task aims to recognize and convert it into natural language. For instance:

- Speech Translation: Converting audio input in source language to text in a target language, requiring the model to understand and accurately translate linguistic nuances.
- Automatic Audio Captioning: Generating descriptive text that captures the sounds and scenes present in the audio, necessitating a comprehensive understanding of the audio context.
- Text-to-Speech Synthesis: Thanks to the neural codec modeling, speech signal can be converted to discrete audio codec, thus generating a sequence based on the target transcription and a segment of prompt speech, a.n.a, zero-shot TTS.

Typically, a neural network learns an end-to-end mapping from audio to text based on a labeled audio-text pair dataset. This is achieved through the following training objective:

$$\mathcal{L}_{\text{att}}(x, y) = -\sum_{l=1}^{L} \log P_\theta(y_l \mid y_{1:l-1}, x) \tag{1}$$

where x is the audio input, y is the corresponding discrete text token sequence, $y_{1:l-1}$ are the tokens generated up to position $l-1$, and θ represents the trainable parameters of the neural model. In the remainder of this chapter, we first introduce the sampling method incorporating the adversarial model design, followed by the corresponding optimization strategy for model θ.

3.1 Model Adversarial Sampling

Given an audio input x with unknown ground truth, two models θ_1 and θ_2 generate their respective text outputs, y_1 and y_2, based on x. An annotator evaluates y_1 and y_2 according to their preferences. If y_1 is preferred, the pair (x, y_1) is added to the set $\mathcal{D}_1^{\text{p}}$, while (x, y_2) is added to $\mathcal{D}_2^{\text{n}}$. Conversely, if y_2 is preferred, (x, y_2) is added to $\mathcal{D}_2^{\text{p}}$ and (x, y_1) to $\mathcal{D}_1^{\text{n}}$. Thus, four sets are created to record the results of the evaluation. Additionally, if both y_1 and y_2 are deemed equally good or bad, they are added to their respective positive and negative sample sets.

Level	BLEU Mean	BLEU Variance
1	40.7	1.10
2	36.6	0.99
3	30.7	0.83
4	28.4	1.21

Table 1: Beam Search BLEU statistics by SeamlessM4T-V2 on De→En translation across different performance level. Each level contains 50 examples and the beam size is 5.

Based on this strategy, m times of human choices can annotate $2 \times m$ data points, which is insufficient for optimizing large models. To encourage more self-generated samples, we use a beam search sampling strategy to expand the preference data. Specifically, the beam search sampling formula is as utilized to generate a set of hypotheses: $\{y^i\}_{i=1}^n = \text{BeamSearch}(x, \theta, n)$, where n is the beam size. The preference label for y^1 will be applied to other candidates $\{y^i\}_{i=2}^n$ in the beam set and stored in the same preference set $\mathcal{D}$, resulting in an n-fold expansion. Although these pseudo-labels inevitably introduce some errors, it is noteworthy that in the beam set generated by audio-to-text models, the candidates do not exhibit the same diversity as those from LLMs. Based on the perceptual module's understanding of the audio input, the generated hypotheses in one beam set exhibit similar quality in terms of metrics.

Empirical validation. We utilize the training set of (Conneau et al. 2023) to show the BLEU diversity in $\{y^i\}_{i=1}^n$ with beam set of 5. As shown in Table 1, each level samples 50 data points according to their average BLEU (± 1), e.g., level 1 consists of 50 set of results with the average BLEU from 40 to 41. Therefore, the variance are calculated on 50 $\times$ 5 utterances for each level. We observe that regardless of SeamlessM4T performance, the variance for each level keeps low, thus indicating the stable BLEU performance in the same beam set. Additionally, similar validations are conducted on other languages to demonstrate that the distribution of BLEU score clusters around the mean value, allowing us to use beam search to provide positive and negative samples with pseudo-preference label.

With proposed model adversarial sampling, we obtain four subsets that totally consists of $m \times n$ of (x, y) pairs. Taking $\mathcal{D}_1^p$ as example, all element are generated by model θ_1 and preferred by annotator, while real label and pseudo label by beam search are recorded to distinguish the data quality. In subsequent optimization, the real label data is assigned higher weight for model update.

Can LLMs help for AST annotation? Given the demonstrated linguistic capabilities of GPT-4, we propose an alternative strategy for preference data annotation in AST. However, since the GPT model lacks audio perception capabilities, this strategy *relies on ground-truth source text* and can only be applied to public AST datasets. Specifically, in the instruction prompt, we provide the transcription of source

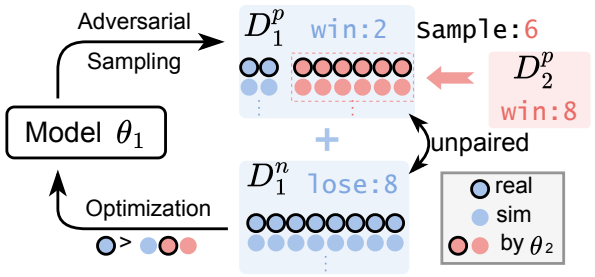

Figure 2: An example of IPO pipeline. In adversarial sampling, model θ_1 wins 2 out of 10 times human evaluation, then the $\mathcal{D}_1^p$ draws further 6 positive samples (red dots) from $\mathcal{D}_2^p$ generated by winning θ_2. During optimization, $\mathcal{D}_1^p$ and $\mathcal{D}_1^n$ provide $8+8$ unpaired positive-negative training examples with different weights: "real" denotes the human-annotated samples, while "sim" refers to those simulated samples generated by beams search.

transcription as a reference and ask LLMs give their preference on 2 models' output.

3.2 Optimization with Imbalanced Data

Given preference dataset $\mathcal{D}$, typical RLHF optimization relies on a reward model r_ϕ learns to serve as a proxy, done by minimizing the negative log-likelihood of the human preference data:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{x,y_w,y_l\sim\mathcal{D}}\left[-\log\sigma\left(r_\phi(x,y_w) - r_\phi(x,y_l)\right)\right] \tag{2}$$

where y_w and y_l denote the preferred and dispreferred data respectively. To avoid excessive pursuit of rewards, a KL divergence penalty is introduced to restrict the language model's deviation from π_{ref}. Here, π_θ represents the model we are optimizing, and the optimal model π^* is defined as the one that maximizes:

$$\mathbb{E}_{x\in\mathcal{D},y\in\pi_\theta}\left[r_\phi(x,y)\right] - \beta D_{\text{KL}}(\pi_\theta(y\mid x)\|\pi_{\text{ref}}(y\mid x)) \tag{3}$$

where $\beta > 0$ is a hyper-parameter. Since this objective function is not differentiable, prior works employ a reinforcement learning algorithm like PPO (Schulman et al. 2017) to optimize it.

However, the performance of PPO-based approach relies on a robust reward model and quite unstable in practice (especially in a distributed setting) (Yuan et al. 2024; Chen et al. 2024c). For this reason, recent work propose a series of closed-form losses that maximize the margin between the preferred and dispreferred generations. As representative algorithms, Direct Preference Optimization (DPO) (Rafailov et al. 2024), has proved its mathematical equivalence with RLHF by minimizing following loss:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta,\pi_{\text{ref}}) = \\ \mathbb{E}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right)\right] \tag{4}$$

However, (4) relies on pairwise preference data that requires distinct (y_w)-(y_l) based on the same input x. Considering the limited diversity in a beam set, we eliminate the dependence on pairwise win-lose samples by utilizing a reference point z_{ref} form recent work KTO (Ethayarajh et al. 2024), and directly maximize the length-normalized implicit reward $r_{\text{IPO}}(x,y)$ with a value function v. Taking model θ_1 as example:

$$v_{\text{IPO}}(x,y_1) = \begin{cases}\sigma(r_{\text{IPO}}(x,y_1) - z_{\text{ref}}) & \text{if } y_1\sim\mathcal{D}_1^p \\ \sigma(z_{\text{ref}} - r_{\text{IPO}}(x,y_1)) & \text{if } y_1\sim\mathcal{D}_1^n\end{cases} \tag{5}$$

$$r_{\text{IPO}}(x,y;\beta) = \frac{\beta}{|y|}\log\frac{\pi_\theta(y\mid x)}{\pi_{\text{ref}}(y\mid x)} \tag{6}$$

where σ is the sigmoid function and the z_{ref} refers to the KL divergence between π_θ and π_{ref}. Furthermore, $|y|$ is the a length penalty that prevent model from generating longer yet low-quality sequence (Meng, Xia, and Chen 2024). In IPO, we simply define it as z_{ref} as $\text{KL}(\pi_\theta(y|x)\parallel\pi_{\text{ref}}(y|x))$, which is calculated in each training batch. Notably, this item is designed to stabilize the training process that is not involved into back-propagation.

To handle the data imbalance issue between each subset, we leverage the samples generated by opposite model to balance the optimization data. As the example shown in Figure 2, model θ_1 outperform θ_2 at 20% of cases, then $\mathcal{D}_1^n$ will be 4 times of $\mathcal{D}_1^p$. In this case, we sample 60% positive data from $\mathcal{D}_2^p$ to extend $\mathcal{D}_1^p$, as the dashed box in Figure 2. Though this parts of data is not generated by model θ_1, maximizing corresponding implicit reward is equivalent to a kind of knowledge distillation from θ_2 on those losing data points of θ_1 . Conversely, the negative examples from $\mathcal{D}_1^n$ also can contribute to extend $\mathcal{D}_2^n$ that prevents model θ_2 from these dispreferred generations. With similar strategy, DPO is reactivated by using non self-generated samples, we select the y_w and y_l in (4) from $\mathcal{D}_1^p$ and $\mathcal{D}_2^n$ since they contains pairwise generations based on the same input x.

Additionally, we utilize the hyper-parameter β in (6) to control the model's updates: when the preference for a given data point comes from human annotators, we set β to 0.1 (same as (Ethayarajh et al. 2024)); however, when it comes from beam search or the adversarial model, β is decreased to 0.05 to regulate the update magnitude. Then the final optimization of IPO can be written as:

$$\mathcal{L}_{\text{IPO}} = \mathbb{E}_{x,y\sim\mathcal{D}}[1 - v_{\text{IPO}}(x,y;\beta)] \tag{7}$$

Due to the calibration of σ in (5), minimizing $\mathcal{L}_{\text{IPO}}$ encourages both models to pursuit higher implicit reward, which guides model to generate preferred samples by annotator.

Generally, IPO can handle *imbalanced* and *unpaired* preference data generated from model adversarial sampling, and effectively improve the performance for both model θ_1 and θ_2. In practice, IPO exhibits two futher advantages: (1) The two models can engage in continuous adversarial-optimization iterations. Although this requires ongoing annotation, when the ground truth of public datasets is available, we demonstrate that LLMs can be used to replace human annotators, as mentioned in Section 3.1. (2) This adversarial optimization method enables efficient distillation, where the preference annotation allows the weak model to selectively absorb knowledge. With just one iteration, the

X→En Model	Ar	Cy	De	El	Es	Fa	Fr	Hi	It	Ja	Pt	Ta	Uk	Vi	Zh	Avg.
Whisper-L-V2	25.5	13.0	34.6	23.7	23.3	19.6	32.2	22.0	23.6	18.9	38.1	9.2	29.4	20.4	18.4	23.5
SeamlessM4T-L-V2	34.7	34.9	37.1	27.3	25.4	30.3	33.7	28.5	26.5	19.5	38.5	22.1	33.2	25.7	23.0	29.4
IPO-Whisper	31.4	19.8	36.5	26.6	24.7	26.5	33.0	27.0	26.0	19.2	38.4	17.5	32.4	24.0	22.5	27.0
+ Iterative Optim.	32.3	20.6	36.8	26.9	25.0	27.5	33.4	27.6	26.0	19.3	38.6	18.3	32.7	24.6	22.8	27.5
w/ Unified Finetune	30.3	18.6	36.2	25.8	24.4	23.6	32.5	25.7	25.2	19.2	38.2	14.6	31.5	23.1	20.8	26.0
IPO-SeamlessM4T	37.6	38.0	40.2	31.1	28.6	32.7	36.6	32.2	28.7	23.3	40.8	26.5	35.4	28.1	26.7	32.4
+ Iterative Optim.	38.5	38.7	41.3	31.5	29.5	33.3	37.2	32.9	29.3	24.0	41.5	27.2	36.3	28.9	27.5	33.2
w/ Unified Finetune	36.0	36.1	39.4	29.8	27.2	30.3	35.6	30.4	28.0	21.9	40.0	24.4	34.5	26.6	25.4	31.0

Table 2: Speech translation results on 15 FLEURS X→En test sets using BLEU score. GPT-4 is used to simulate the 400 annotations for each round, with given source-language and target ground-truth text to make full use of its multilingual ability. "Iterative Optim." denotes the result that repeat sampling-learning iterations for 3 times, and "Unified Finetune" denotes train one model using across all languages.

# Annotations	Whisper / SeamlessM4T		
	De→En	Fr→En	Zh→En
0	34.6 / 37.1	32.2 / 33.7	18.4 / 23.0
100	35.7 / 38.2	33.9 / 34.7	21.7 / 24.9
200	36.6 / 40.2	33.3 / 36.8	22.5 / 26.8
400	**36.9 / 40.6**	**33.5 / 37.2**	**22.9 / 27.1**
400 *w/o B.S.*	35.9 / 38.8	33.6 / 34.4	21.9 / 25.9

Table 3: Speech translation results on FLEURS X→En test sets using BLEU score. Human evaluators are incorporated to annotate data. "*w/o B.S.*" denotes the ablation results without pseudo-preference by beam search.

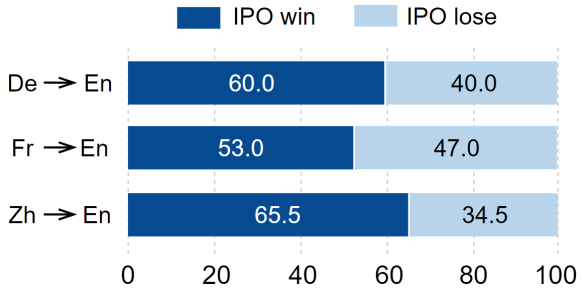

Figure 3: Subjective A/B test of IPO and baseline on AST. Whisper and SeamlessM4T are compared with the baseline separately, and then their win and loss rates are combined for presentation according to translated languages.

weak model (can be a small model) significantly improve its performance, and it does not require the two models to have identical vocabularies.

4 Results on Speech Translation

4.1 Baseline Model and Experimental Setup

In this work, we select two popular large speech-to-text models as our starting point, i.e., SeamlessM4T-large-v2 (Barrault et al. 2023) and Whisper-large-v2 (Radford et al. 2023). SeamlessM4T is tailored to full-modality multilingual translation tasks and has achieved the state-of-the-art on various AST benchmarks and language directions. Whisper is tailored to robust speech recognition and also shows good robustness on X→En speech translation tracks.

For experiments, we use the FLEURS (Conneau et al. 2023) benchmark to evaluate our proposed approach, which is one of the most popular AST benchmarks. Specifically, we select 15 common X→En language directions in this study. For sampling, we use the two large models introduced above with beam size of 5. For preference data annotation, we employ both human listener and GPT-4 as we introduced in Section 3.1. Considering the high cost of searching for native speakers in various languages, three major languages are selected for real human annotation–German, French, and Chinese. These 3 evaluators are all X-language native speakers while with excellent English level (live in English-speaking region more than 3 years). After listening source language speech, they compare the output results from two models and then give their preference. Then, we also investigate GPT-4 to annotate data in all 15 languages, whose superior multilingual ability enables an efficient and automatic annotation pipeline. Since we have access to the source-language ground-truth text in AST benchmark, we provide it to GPT-4 to obtain more comprehensive annotations. Furthermore, we investigate two settings in experiments, i.e., *specific finetuning* and *unified finetuning*. The former indicates that we annotate data and finetune foundation model on each source language individually, while the latter indicates that we do everything for all source languages together and only obtain all-in-one finetune model at last, we will report their results respectively. Thereafter, for evaluation we use BLEU score with ground-truth, and we also invite human evaluators to listen to source speech and evaluate the translated text according their preference.

4.2 Objective Evaluation

Table 3 illustrates the AST results on three FLEURS X→En directions with both Whisper and SeamlessM4T baselines. We can observe that only 100 annotations can produce clear BLEU improvements over two strong baseline models and three language directions, e.g., 1.1 BLEU improvement on SeamlessM4T De→En, which demonstrates the remarkable data efficiency of our approach. Thereafter, with increasing data annotation, we can obtain better BLEU results on different baseline models and translations, showing the perfor-

mance gain obtained from human feedback. The last row also demonstrates the efficacy with beam search by ablation.

Table 2 illustrates the AST results on 15 FLEURS X→En directions with GPT-4 as the preference data annotator. We first observe from the two baselines that Whisper-large-v2 achieves satisfactory performance (as a robust speech recognition model) on all source languages except Tamil, while SeamlessM4T-large-v2 achieves state-the-of-art performance on all languages. Therefore, in this study, our major goal is to improve the Whisper model with guidance from SeamlessM4T. As expected, our proposed IPO significantly improve the performance of Whisper baseline model, especially the almost 10 BLEU increase on Ta→En track that vanilla Whisper is not good at. The overall gain is 3.5 BLEU score, indicating the effectiveness of IPO in improving weak baseline model with adversarial sampling and RLHF finetuning. On the other hand, we are surprising to find that the stronger baseline model, SeamlessM4T, also obtained great enhancement with a gain of 3.0 BLEU score. This finding indicates that our RLHF finetuning approach with adversarial sampling enables different-level models to benefit from each other, resulting in a new state-of-the-art based on the strong SeamlessM4T baseline model.

In addition, we also conduct two studies to investigate *iterative optimization* and *unified finetuning*. (1) The proposed sampling-learning pipeline can be iteratively conducted to further improve the model performance, as updated model can yield better sampling results. We conduct 3 times of iterative finetuning on top of previously optimized model and present the results as "Iterative Optim." in Table 2, which achieves even further improvements over IPO results. However, more iterative rounds would not obtain more performance gain, approaching the upper-bound performance limited preference data. (2) We also perform unified finetune using all data together across multiple languages, i.e., "Unified Finetune" in Table 2. Although slightly inferior to specific fine-tuning, it is still significantly better than the Whisper and SeamlessM4T baseline. More importantly, unified finetuning achieves better generalization ability without forgetting on specific language.

4.3 Subjective Evaluation

For comprehensive assessment, we also incorporate the human evaluators for subjective evaluation. For each language, 100 examples are randomly selected from test set. Whisper and SeamlessM4T (after IPO) respectively perform inference, resulting in 2 $\times$ 100 translation utterances. Then the A/B test is conducted on the comparison between theses utterances with baseline (before IPO). As shown in Fig. 3, our IPO shows consistent and clear advantage over baseline on three language translations, which verifies the effectiveness of our proposed approach.

5 Results on Audio Captioning

5.1 Baseline Model and Experimental Setup

Automatic audio captioning (AAC) is a crucial task in understanding real-world audio signals, which requires a sound description of the events happening in given audio.

# Annotations	Pengi / Qwen-Audio CIDEr ↑	SPICE ↑	SPIDEr ↑
0	0.416 / 0.441	0.126 / 0.136	0.271 / 0.288
100	0.435 / 0.460	0.130 / 0.139	0.283 / 0.299
200	0.444 / 0.469	0.132 / 0.141	0.288 / 0.302
400	**0.448 / 0.473**	**0.133 / 0.141**	**0.291 / 0.306**
400 *w/o B.S.*	0.440 / 0.447	0.130 / 0.138	0.285 / 0.294

Table 4: Audio captioning results on Clotho dataset. "*w/o B.S.*" denotes the ablation results without pseudo-preference by beam search. Pengi win rate: 41.0%

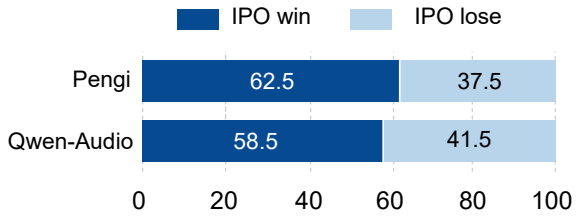

Figure 4: Subjective A/B test of proposed IPO and baseline on AAC. Human evaluator are included to give preference.

In this work, we select two large audio language models that can perform audio captioning as our starting point, i.e., Pengi (Deshmukh et al. 2023) and Qwen-Audio (Chu et al. 2023), both of which attempt to incorporate audio modality into language model for comprehensive understanding.

For experiments, we the Clotho (Drossos, Lipping, and Virtanen 2020) benchmark to evaluate our proposed approach, which is among the most popular AAC benchmarks. For adversarial sampling of IPO, two human listener are employ to give total 400 annotations after listening audio segments. It is noted that human annotation prefer longer captions when comparing the output of two models, therefore, the length penalty is important to prevent model from generating too long sentences. For evaluation, we use three metrics, i.e., CIDEr, SPICE and SPIDEr, to compare our optimized model with baseline. In A/B test, human evaluators are required to give the preference on 200 captioning results between before/after IPO.

5.2 Objective and Subjective Evaluation

Table 4 illustrates the AAC results on two baselines with different number of annotations in training data. First, we can observe that the two baseline models achieve quite good performance on Clotho benchmark. However, in our trials, only 100 annotations can yield significant improvements on both baseline models in terms of three metrics, especially the CIDEr. Increasing annotations can help further improve the performance, where 400 seems to approach the upper-bound. Overall, we can observe that IPO not only significantly improves the weak baseline (i.e., Pengi), but it also enhances the stronger one (i.e., Qwen-Audio), similar to what we observed previously in AST experiments. Furthermore, subjective A/B test in Fig. 4 also shows the effectiveness of our proposed IPO.

Method	Predicted Captioning Result
Pengi	People talk with cars driving by
Qwen-Audio	A men is talking with busy traffic and sharp noise in the background
IPO-Pengi	People talk with drilling noise and busy traffic behind
IPO-Qwen-Audio	A men talks with jackhammer running and busy traffic roaring by in the background

Table 5: Case study of audio captioning on Clotho dataset. Model provide more detailed description for audio event via IPO.

# Annotations	VC-330M / VC-830M MOS ↑	WER ↓	SIM ↑
0	3.37 / 4.21	17.9 / 2.7	0.76 / 0.90
200	3.85 / 4.35	8.0 / 2.5	0.87 / 0.92

Table 6: Objective Zero-shot TTS results on LibriTTS dataset. "MOS", "WER", and "SIM" are all calculated by pre-trained nerual model. VC-330M win rate: 18.5%.

Case Study. To demonstrate the enhancement of the model's perception abilities by IPO, we provide a case study in Table 5. The two baselines both give a description about people talking and the cars driving by, where the stronger Qwen-Audio specifies more details about the male and sharp noise. In comparison, IPO optimized models yield more details, e.g., drilling noise, jackhammer running, which more comprehensively describe the audio event.

6 Extension on Speech Syntheses

6.1 Baseline Model and Experimental Setup

To comprehensive evaluate the proposed IPO approach, we also conduct experiments on zero-shot text-to-speech (TTS) synthesis task. For experiments, we select two size of zero-shot TTS system which are proposed in VoiceCraft (Peng et al. 2024). i.e., VC-330M and VC-830M as our baseline model, and we select the popular LibriTTS dataset (Zen et al. 2019) as benchmark. Notably, this experiment can be also demonstrate that IPO can be utilized as a knowledge distillation function from large model to light model. We does not use beam search but perform inference for 5 times with random seeds. For evaluation, we employ NISQA (Mittag et al. 2021) to calculate mean opinion score (MOS), Whisper-medium.en to calculate word error rate, and WavLM-TDCNN to calculate speaker similarity (SIM).

During sampling, 2 human listeners are employed to respectively give 100 annotations for two models, resulting in 200 samples that VC-830M win for 163 times. Despite the disparity in results, we find that VC-330M mostly lose due to its robustness on some unseen speakers, since its training data does not include LibriTTS. For evaluation, these 2 human listeners perform A/B test on 100 examples (6 second to 16 second) sampled from LirbriTTS test set it terms of naturalness of synthesized speech, including a tie if it is hard to give their preference.

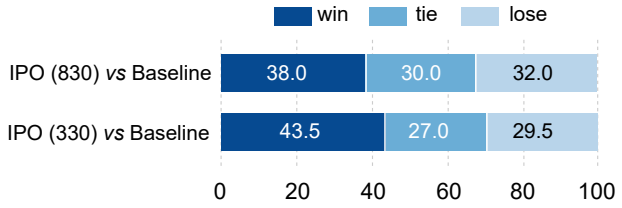

Figure 5: Subjective A/B test of proposed IPO and baseline on TTS naturalness.

6.2 Objective Evaluation

Table 6 illustrates the objective evaluation results in terms of MOS, WER, and SIM. The experimental results on IPO indicate that this 200 annotations effectively improve the synthesized speech quality in terms of three metrics. Specifically, the VC-330M has been improved to a large extent, the failed cases are significantly avoided according to the ratio of high WER>50% samples (10.4%→3.1%). The VC-830M baseline can even be further enhanced, which is similar to AST and AAC tasks. This experimental evidence on TTS demonstrates the remarkable scalability of IPO.

6.3 Subjective Evaluation

Furthermore, we present some subjective evaluation in Fig. 5 to verify the effectiveness of our approach. We conduct A/B test between our optimized model and VC-330M baseline and ground-truth speech by asking human evaluators to select the better one. Results indicate that our synthesized speech outperforms the baseline (i.e., larger win rate than lose rate), though it still lags a bit behind the ground-truth speech. It indicates that our approach makes good use of various baseline models to sample and annotate data, which is proved beneficial to both models.

7 Conclusion

In this work, we explore a novel topic that align audio-text generation with subjective feedback. To this end, we propose IPO, consisting of a model adversarial sampling strategy to alleviate the limited annotations, and a interactively optimization algorithm based on imbalanced and unpaired preference data. Experimental evidence demonstrates the significant performance gain of IPO on three cross-modal generation tasks including AST, AAC, and TTS, enhancing the performance of multiple prevalent audio-text pre-trained models. More importantly, our approach provides new insight and inspiration to the post-training alignment of these models using limited human annotations.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.-A.; Ellis, B.; Elsahar, H.; Haaheim, J.; et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*.

Chen, C.; Hu, Y.; Wu, W.; Wang, H.; Chng, E. S.; and Zhang, C. 2024a. Enhancing Zero-shot Text-to-Speech Synthesis with Human Feedback. *arXiv preprint arXiv:2406.00654*.

Chen, X.; Zhang, S.; Bai, Q.; Chen, K.; and Nakamura, S. 2024b. LLaST: Improved End-to-end Speech Translation System Leveraged by Large Language Models. *arXiv preprint arXiv:2407.15415*.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024c. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Cheng, P.; Yang, Y.; Li, J.; Dai, Y.; and Du, N. 2023. Adversarial preference optimization. *arXiv preprint arXiv:2311.08045*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Chuang, Y.-S.; Liu, C.-L.; Lee, H.-Y.; and Lee, L.-s. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*.

Cideron, G.; Girgin, S.; Verzetti, M.; Vincent, D.; Kastelic, M.; Borsos, Z.; McWilliams, B.; Ungureanu, V.; Bachem, O.; Pietquin, O.; et al. 2024. Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*.

Conneau, A.; Ma, M.; Khanuja, S.; Zhang, Y.; Axelrod, V.; Dalmia, S.; Riesa, J.; Rivera, C.; and Bapna, A. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 798–805. IEEE.

Deshmukh, S.; Elizalde, B.; Singh, R.; and Wang, H. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36: 18090–18108.

Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 736–740. IEEE.

Dubois, Y.; Li, C. X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P. S.; and Hashimoto, T. B. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Fathullah, Y.; Wu, C.; Lakomkin, E.; Li, K.; Jia, J.; Shangguan, Y.; Mahadeokar, J.; Kalinli, O.; Fuegen, C.; and Seltzer, M. 2024. AudioChatLlama: Towards General-Purpose Speech Abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5522–5532.

Gibbons, A.; Donohue, I.; Gorman, C.; King, E.; and Parnell, A. 2023. NEAL: an open-source tool for audio annotation. *PeerJ*, 11: e15913.

He, Z.; Wang, X.; Jiao, W.; Zhang, Z.; Wang, R.; Shi, S.; and Tu, Z. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*.

Hu, S.; Zhou, L.; Liu, S.; Chen, S.; Hao, H.; Pan, J.; Liu, X.; Li, J.; Sivasankaran, S.; Liu, L.; et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, 13916–13932. PMLR.

Kaur, N.; and Singh, P. 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7): 5837–5880.

Kim, Y.; Jang, J.; and Shin, S. 2022. Music2Video: Automatic Generation of Music Video with fusion of audio and text. *arXiv preprint arXiv:2201.03809*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, D.; Tang, C.; and Liu, H. 2024. Audio-LLM: Activating the Capabilities of Large Language Models to Comprehend Audio Data. In *International Symposium on Neural Networks*, 133–142. Springer.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19401–19411.

Liu, R.; Xu, G.; Jia, C.; Ma, W.; Wang, L.; and Vosoughi, S. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. *arXiv preprint arXiv:2012.02952*.

MacGlashan, J.; Ho, M. K.; Loftin, R.; Peng, B.; Wang, G.; Roberts, D. L.; Taylor, M. E.; and Littman, M. L. 2017. Interactive learning from policy-dependent human feedback.

In *International conference on machine learning*, 2285–2294. PMLR.

Majumder, N.; Hung, C.-Y.; Ghosal, D.; Hsu, W.-N.; Mihalcea, R.; and Poria, S. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *arXiv preprint arXiv:2404.09956*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

Mousavi, S. M.; Roccabruna, G.; Alghisi, S.; Rizzoli, M.; Ravanelli, M.; and Riccardi, G. 2024. Are LLMs Robust for Spoken Dialogues? *arXiv preprint arXiv:2401.02297*.

Peng, P.; Huang, P.-Y.; Li, D.; Mohamed, A.; and Harwath, D. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. *arXiv preprint arXiv:2403.16973*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Rubenstein, P. K.; Asawaroengchai, C.; Nguyen, D. D.; Bapna, A.; Borsos, Z.; Quitry, F. d. C.; Chen, P.; Badawy, D. E.; Han, W.; Kharitonov, E.; et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Schneider, F.; Kamal, O.; Jin, Z.; and Schölkopf, B. 2023. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Steenhoek, B.; Tufano, M.; Sundaresan, N.; and Svyatkovskiy, A. 2023. Reinforcement learning from automatic feedback for high-quality unit test generation. *arXiv preprint arXiv:2310.02368*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Sun, Z.; Shen, Y.; Zhang, H.; Zhou, Q.; Chen, Z.; Cox, D. D.; Yang, Y.; and Gan, C. 2024. SALMON: Self-Alignment with Instructable Reward Models. In *The Twelfth International Conference on Learning Representations*.

Verma, A.; Agarwal, S.; Arya, K.; Petrlik, I.; Esparza, R.; and Rodriguez, C. 2023. Image Captioning with Reinforcement Learning. In *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 1–7. IEEE.

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. F. 2024. AudioBench: A Universal Benchmark for Audio Large Language Models. *arXiv preprint arXiv:2406.16020*.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2096–2105.

Xu, L.; Tang, Q.; Lv, J.; Zheng, B.; Zeng, X.; and Li, W. 2023. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546: 126287.

Xu, N.; Zhao, J.; Zu, C.; Gui, T.; Zhang, Q.; and Huang, X. 2024. Advancing Translation Preference Modeling with RLHF: A Step Towards Cost-Effective Solution. *arXiv preprint arXiv:2402.11525*.

Yang, G.; Ma, Z.; Yu, F.; Gao, Z.; Zhang, S.; and Chen, X. 2024. MaLa-ASR: Multimedia-Assisted LLM-Based ASR. *arXiv preprint arXiv:2406.05839*.

Yuan, W.; Pang, R. Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Zhao, Z.; Jiang, Y.; Liu, H.; Wang, Y.; and Wang, Y. 2024. LibriSQA: A Novel Dataset and Framework for Spoken Question Answering with Large Language Models. *IEEE Transactions on Artificial Intelligence*.

Zhou, P.; Wang, L.; Liu, Z.; Hao, Y.; Hui, P.; Tarkoma, S.; and Kangasharju, J. 2024. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*.